

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

Université de Monastir

**_*_*_

Institut Supérieure d'Informatique et Mathématique de Monastir



Mini Projet

**Techniques d'analyse de données-Concepts et techniques de
fouilles de données**

Réalisé par

Fahed ABDELLAOUI

Année Universitaire : 2015/2016

Projet génie des connaissances

Techniques d'analyse de données-Concepts et techniques de fouilles de données

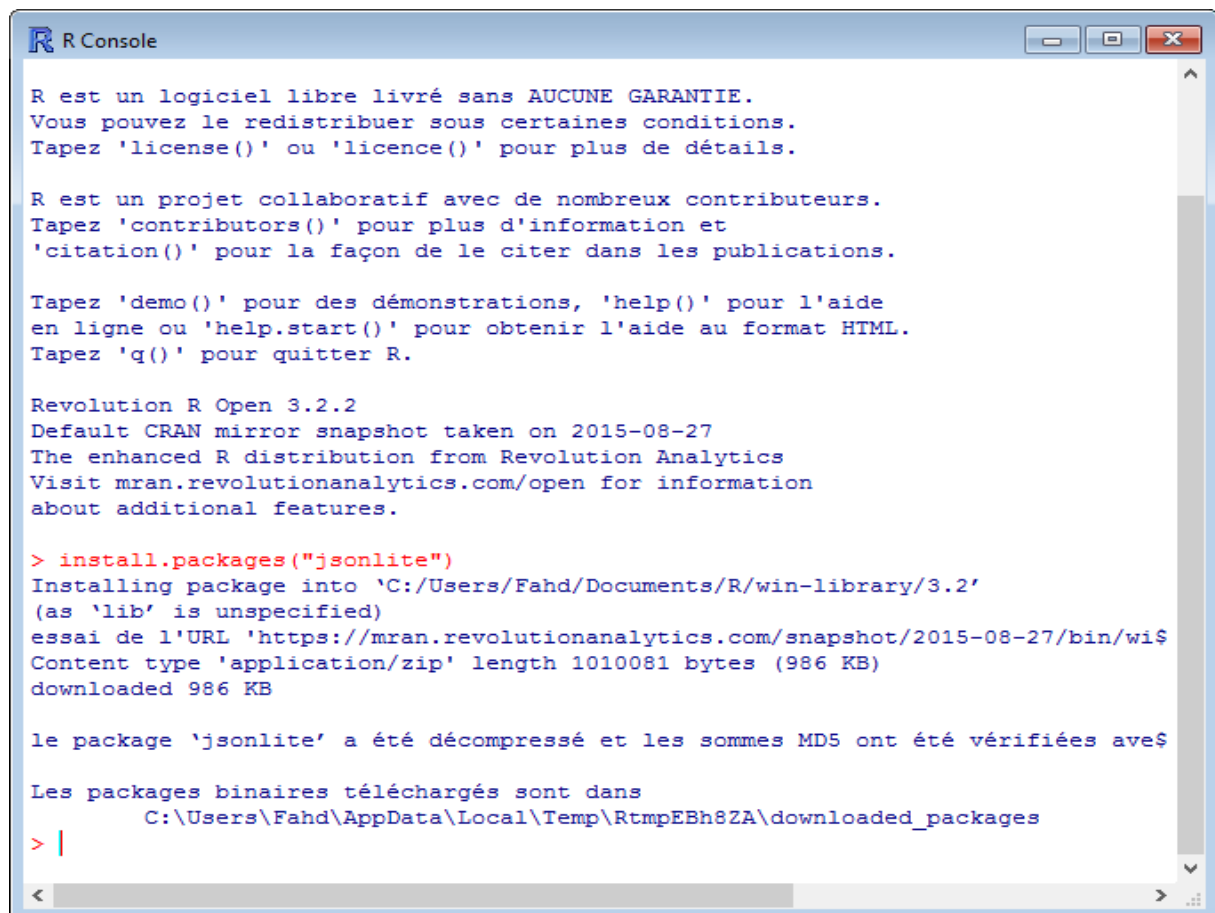
Objectif du projet : analytique des données de l'enseignement supérieur

Base De Données : fr-esr-insertion_professionnelle-master_donnees_nationales.json

Etape 1 : Importation de la base de données (.json) et conversion au type data frame :

L'importation d'une base de données au format JSON dans le logiciel R nécessite les fonctionnalités du package « jsonlite ». Le package jsonlite est un convertisseur / générateur de JSON optimisé. Sa principale force est qu'il implémente une correspondance bidirectionnelle entre les données JSON et les types les plus importants de données de logiciel R. Ainsi nous pouvons convertir entre les objets R et JSON sans perte de type ou d'information, et sans la nécessité de manuels des données. Ceci est idéal pour interagir avec bases des données avec des nombreux enregistrements et pour les explorer de manière transparente sur R.

Pour Bien importer la base de données sous format JSON on doit charger le package « jsonlite ». Les étapes suivantes représentent l'importation de ce package dans le logiciel R :



```
R Console

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

Revolution R Open 3.2.2
Default CRAN mirror snapshot taken on 2015-08-27
The enhanced R distribution from Revolution Analytics
Visit mran.revolutionanalytics.com/open for information
about additional features.

> install.packages("jsonlite")
Installing package into 'C:/Users/Fahd/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
essai de l'URL 'https://mran.revolutionanalytics.com/snapshot/2015-08-27/bin/wi$
Content type 'application/zip' length 1010081 bytes (986 KB)
downloaded 986 KB

le package 'jsonlite' a été décompressé et les sommes MD5 ont été vérifiées ave$

Les packages binaires téléchargés sont dans
  C:\Users\Fahd\AppData\Local\Temp\RtmpEBh8ZA\downloaded_packages
> |
```

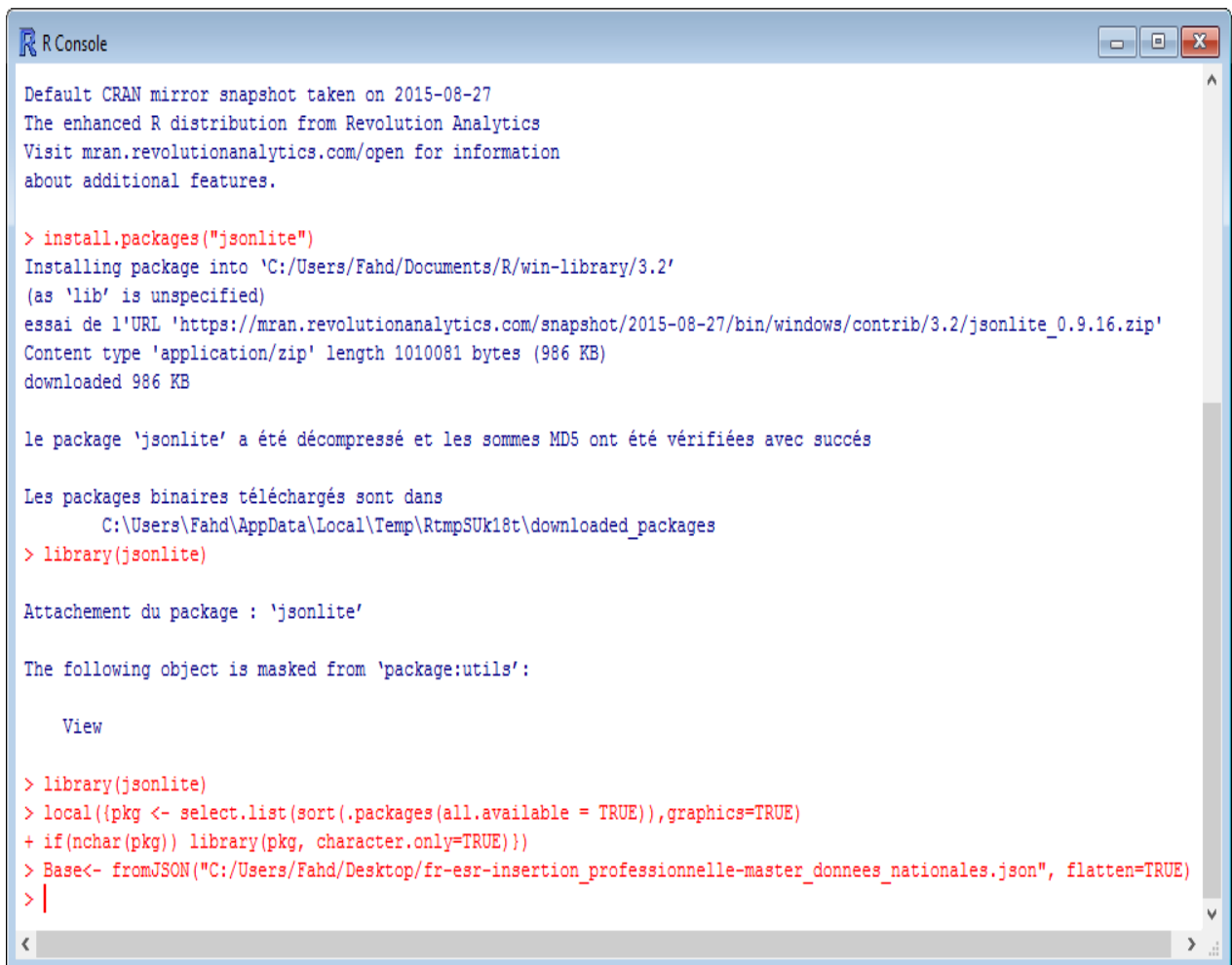
Figure 1 : Installation du package 'jsonlite'

Après le téléchargement et l'installation du package « jsonlite » on doit l'importer pour l'utiliser, pour cela on doit le sélectionner à partir de l'onglet Package > Charger le package puis on tape : library(jsonlite). Maintenant le package est prêt à être utilisé.

Si notre fichier JSON qui comporte la base de données est dans le répertoire courant on doit taper : getwd() pour copier les fichiers dans le répertoire courant, puis on tape :

Dans notre cas le fichier JSON est nommé : fr-esr-insertion_professionnelle-master_donnees_nationales.json on doit taper : Base<- fromJSON("fr-esr-insertion_professionnelle-master_donnees_nationales.json", flatten=TRUE).

Si le fichier est dans un autre emplacement on doit spécifier le chemin dans lequel il se trouve.



```
R Console
Default CRAN mirror snapshot taken on 2015-08-27
The enhanced R distribution from Revolution Analytics
Visit mran.revolutionanalytics.com/open for information
about additional features.

> install.packages("jsonlite")
Installing package into 'C:/Users/Fahd/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
essai de l'URL 'https://mran.revolutionanalytics.com/snapshot/2015-08-27/bin/windows/contrib/3.2/jsonlite_0.9.16.zip'
Content type 'application/zip' length 1010081 bytes (986 KB)
downloaded 986 KB

le package 'jsonlite' a été décompressé et les sommes MD5 ont été vérifiées avec succès

Les packages binaires téléchargés sont dans
C:\Users\Fahd\AppData\Local\Temp\RtmpSUK18t\downloaded_packages
> library(jsonlite)

Attachement du package : 'jsonlite'

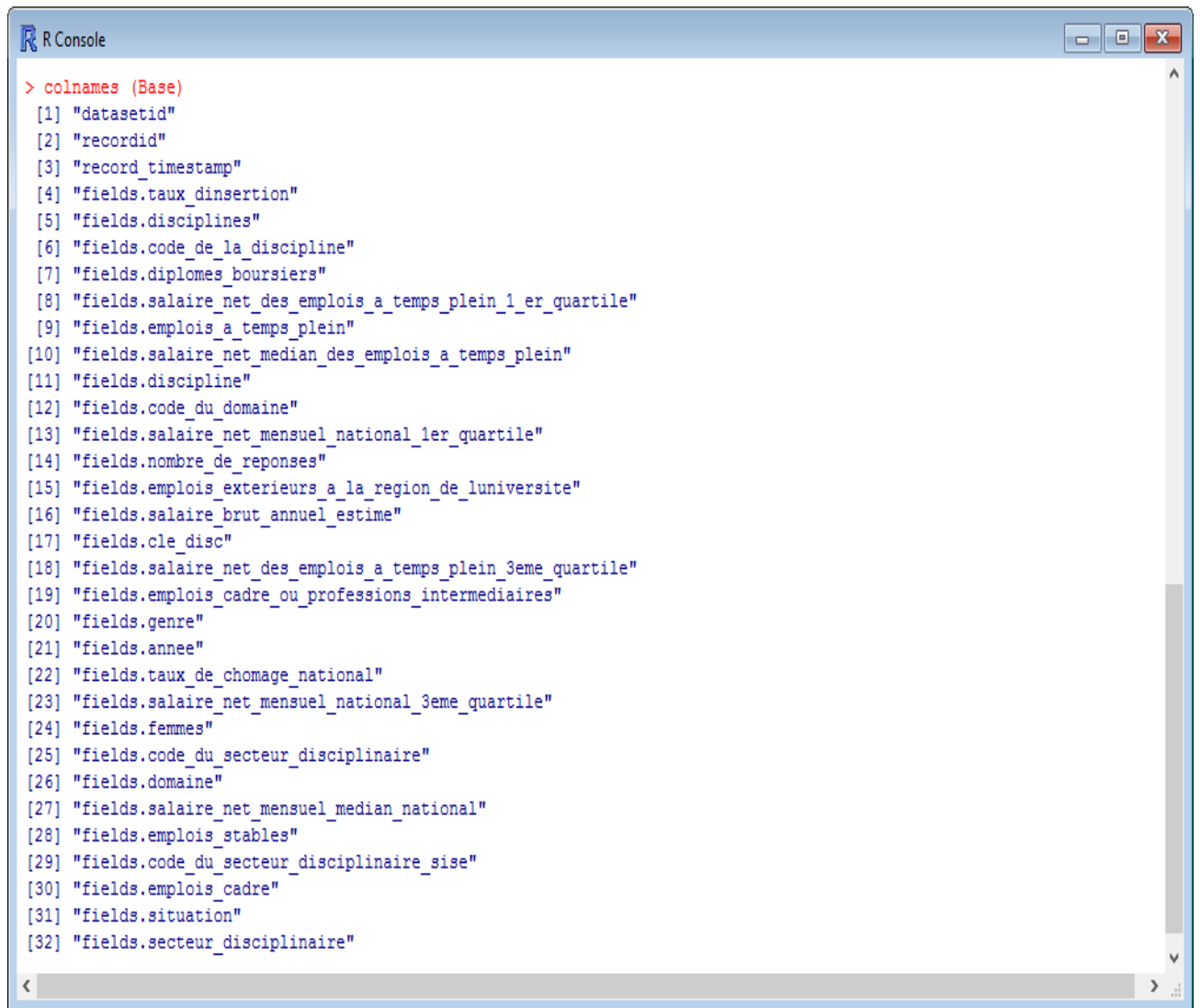
The following object is masked from 'package:utils':

    View

> library(jsonlite)
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
> Base<- fromJSON("C:/Users/Fahd/Desktop/fr-esr-insertion_professionnelle-master_donnees_nationales.json", flatten=TRUE)
> |
```

Figure 2 : Importation du fichier JSON

Maintenant la base est importé vers le format data. Frame et on peut l'utilisé en tant que R Object. Pour Lister les noms de colonnes de la base on peut taper : colnames(Base).



```
> colnames (Base)
[1] "datasetid"
[2] "recordid"
[3] "record_timestamp"
[4] "fields.taux_dinsertion"
[5] "fields.disciplines"
[6] "fields.code_de_la_discipline"
[7] "fields.diplomes_boursiers"
[8] "fields.salaire_net_des_emplois_a_temps_plein_1er_quartile"
[9] "fields.emplois_a_temps_plein"
[10] "fields.salaire_net_median_des_emplois_a_temps_plein"
[11] "fields.discipline"
[12] "fields.code_du_domaine"
[13] "fields.salaire_net_mensuel_national_1er_quartile"
[14] "fields.nombre_de_reponses"
[15] "fields.emplois_exterieurs_a_la_region_de_luniversite"
[16] "fields.salaire_brut_annuel_estime"
[17] "fields.cle_disc"
[18] "fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile"
[19] "fields.emplois_cadre_ou_professions_intermediaires"
[20] "fields.genre"
[21] "fields.annee"
[22] "fields.taux_de_chomage_national"
[23] "fields.salaire_net_mensuel_national_3eme_quartile"
[24] "fields.femmes"
[25] "fields.code_du_secteur_disciplinaire"
[26] "fields.domaine"
[27] "fields.salaire_net_mensuel_median_national"
[28] "fields.emplois_stables"
[29] "fields.code_du_secteur_disciplinaire_sise"
[30] "fields.emplois_cadre"
[31] "fields.situation"
[32] "fields.secteur_disciplinaire"
```

Figure 3 : Affichage nom des colonnes de la base

Comme la montre la figure 3, La base comporte 32 colonnes le but de ce travail est de bien explorer la base et d'étudier les enregistrements.

On a fini l'étape 1 qui consiste à bien importer la base des données au format JSON et la transformer à un data. Frame.

Etape 2 : Description de la base de données :

La Description de la base de données consiste à lister :

- Nombre d'objets.
- Nombres d'attributs.
- les attributs et leurs types.
- Comment les données ont été collectées ?
- Description de chaque attribut.
- Les valeurs possibles de chaque attribut.

1. Nombre d'objets et d'attributs :

Comme la montre la figure ci-dessous (Figure 4) la base comporte :

- 389 Objets (Enregistrement).
- 32 attributs



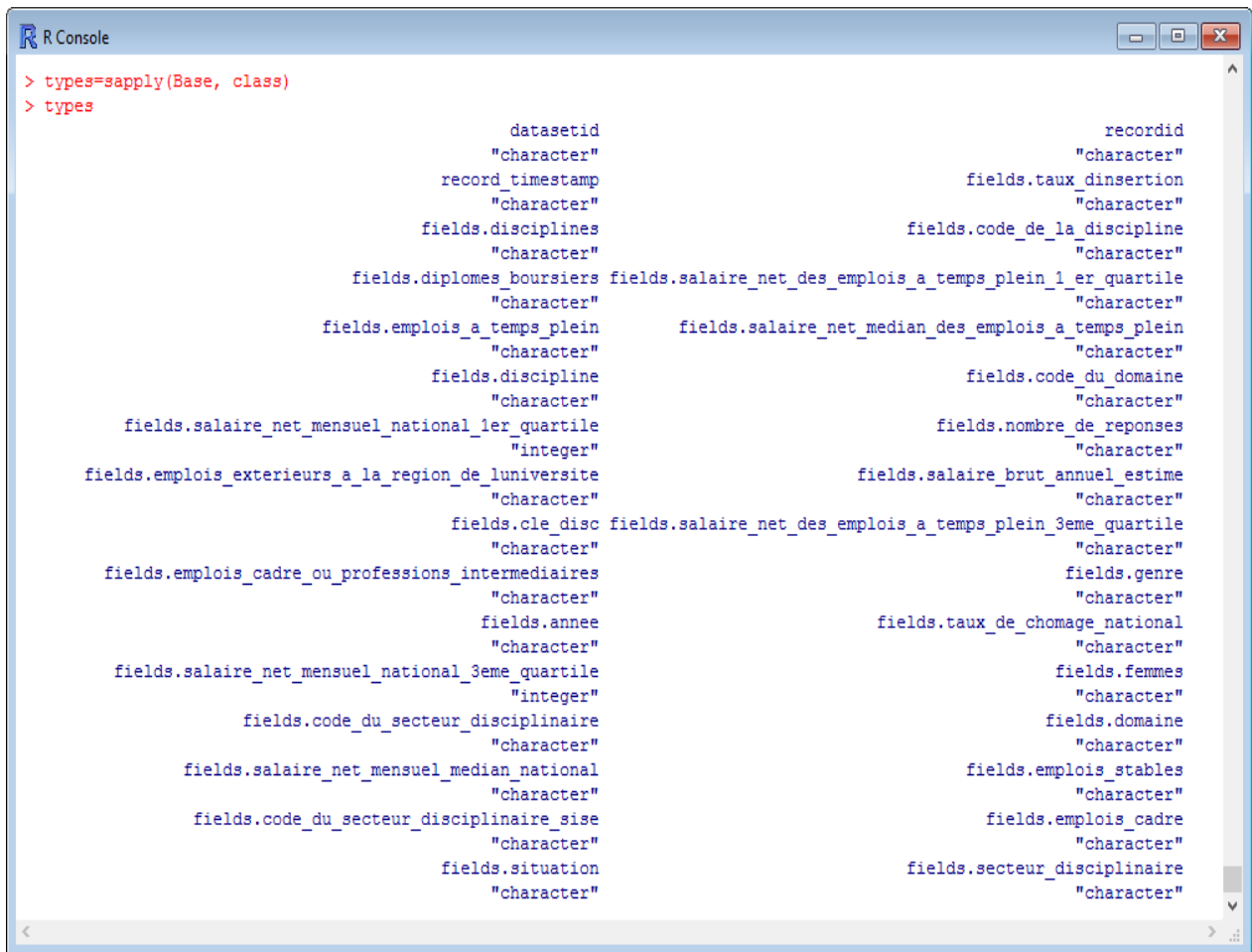
```
R Console
[3] "record_timestamp"
[4] "fields.taux_dinsertion"
[5] "fields.disciplines"
[6] "fields.code_de_la_discipline"
[7] "fields.diplomes_boursiers"
[8] "fields.salaire_net_des_emplois_a_temps_plein_1er_quartile"
[9] "fields.emplois_a_temps_plein"
[10] "fields.salaire_net_median_des_emplois_a_temps_plein"
[11] "fields.discipline"
[12] "fields.code_du_domaine"
[13] "fields.salaire_net_mensuel_national_1er_quartile"
[14] "fields.nombre_de_reponses"
[15] "fields.emplois_exterieurs_a_la_region_de_luniversite"
[16] "fields.salaire_brut_annuel_estime"
[17] "fields.cle_disc"
[18] "fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile"
[19] "fields.emplois_cadre_ou_professions_intermediaires"
[20] "fields.genre"
[21] "fields.annee"
[22] "fields.taux_de_chomage_national"
[23] "fields.salaire_net_mensuel_national_3eme_quartile"
[24] "fields.femmes"
[25] "fields.code_du_secteur_disciplinaire"
[26] "fields.domaine"
[27] "fields.salaire_net_mensuel_median_national"
[28] "fields.emplois_stables"
[29] "fields.code_du_secteur_disciplinaire_sise"
[30] "fields.emplois_cadre"
[31] "fields.situation"
[32] "fields.secteur_disciplinaire"
> dim(Base)
[1] 389 32
> |
```

Figure 4 : Affichage dimensions de la base

2. Attributs et types :

Pour déterminer le type de chaque attribut on peut taper : `sapply (Base, class)`

Le résultat est représenté par la figure ci-dessous (Figure 5) :



```
R Console
> types=sapply(Base, class)
> types
datasetid                                recordid
"character"                             "character"
record_timestamp                        fields.taux_dinsertion
"character"                             "character"
fields.disciplines                      fields.code_de_la_discipline
"character"                             "character"
fields.diplomes_boursiers              fields.salaire_net_des_emplois_a_temps_plein_1_er_quartile
"character"                             "character"
fields.emplois_a_temps_plein            fields.salaire_net_median_des_emplois_a_temps_plein
"character"                             "character"
fields.discipline                      fields.code_du_domaine
"character"                             "character"
fields.salaire_net_mensuel_national_1er_quartile
fields.emplois_exterieurs_a_la_region_de_luniversite
"integer"                               fields.nombre_de_reponses
"character"                             "character"
fields.cle_disc                        fields.salaire_brut_annuel_estime
"character"                             "character"
fields.emplois_cadre_ou_professions_intermediaires
fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile
"character"                             "character"
fields.annee                          fields.genre
"character"                             "character"
fields.salaire_net_mensuel_national_3eme_quartile
fields.taux_de_chomage_national
"integer"                               "character"
fields.code_du_secteur_disciplinaire    fields.femmes
"character"                             "character"
fields.salaire_net_mensuel_median_national
fields.domaine
"character"                             "character"
fields.code_du_secteur_disciplinaire_size
fields.emplois_stables
"character"                             "character"
fields.situation                      fields.emplois_cadre
"character"                             "character"
fields.secteur_disciplinaire
"character"                             "character"
```

Figure 5 : Type des attributs

Remarque :

Cela ne peut pas nous donner une idée claire sur le type exacte des attributs il faut explorer la base pour bien extraire les types des attributs :
(Nominal,Ordinal,Binary,Interval,Ratio,Discrete,Continious...)

Ce tableau représente l'ensemble des attributs de la base, les éléments en rouge sont les éléments pertinentes.

Numéro de l'attribut	Nom de l'attribut	Type de l'attribut
1	datasetid	Nominal
2	recordid	Nominal
3	record_timestamp	Interval
4	fields.taux_dinsertion	Ratio
5	fields.disciplines	Ordinal
6	fields.code_de_la_discipline	Nominal
7	fields.diplomes_boursiers	Ratio
8	fields.salaire_net_des_emplois_a_temps_plein_1er_quartile	Continuous Attribute
9	fields.emplois_a_temps_plein	Ratio
10	fields.salaire_net_median_des_emplois_a_temps_plein	Continuous Attribute
11	fields.discipline	Ordinal
12	fields.code_du_domaine	Nominal
13	fields.salaire_net_mensuel_national_1er_quartile	Continuous Attribute
14	fields.nombre_de_reponses	Ratio
15	fields.emplois_exterieurs_a_la_region_de_luniversite	Ratio
16	fields.salaire_brut_annuel_estime	Continuous Attribute
17	fields.cle_disc	Nominal
18	fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile	Continuous Attribute
19	fields.emplois_cadre_ou_professions_intermediaires	Nominal
20	fields.genre	Binary
21	fields.annee	Interval
22	fields.taux_de_chomage_national	Continuous Attribute
23	fields.salaire_net_mensuel_national_3eme_quartile	Continuous Attribute
24	fields.femmes	Ratio
25	fields.code_du_secteur_disciplinaire	Nominal
26	fields.domaine	Ordinal
27	fields.salaire_net_mensuel_median_national	Continuous Attribute
28	fields.emplois_stables	Ratio
29	fields.code_du_secteur_disciplinaire_sise	Nominal
30	fields.emplois_cadre	Ratio
31	fields.situation	Interval
32	fields.secteur_disciplinaire	Ordinal

Les attributs pertinents seront utilisés ultérieurement pour analyser la base de données, les autres attributs ne seront pas pris en compte car ils manquent d'importance ou un manque grave des valeurs (données manquantes).

3. Comment les données ont été collectées ?

Cette base de données permet la représentation des données relatives aux étudiants qui ont des diplômes de Master dans les différents domaines, Cette base donne une idée sur l'emploi de ces étudiants en extérieur de la région de leurs études universitaire, le calcul de leurs salaire, leurs conditions : temps de chômage avant le recrutement, le taux de chômage national, le classement selon le sexe et d'autres informations supplémentaires.

Les données sont relatives aux années 2011/2012 mais elles sont collectées en 2015.

4. Description des attributs :

1. **Datasetid** : Comporte l'identificateur de la base des données, cette valeur est identique pour tous les enregistrements de cette base de données.
2. **Recordid** : Représente l'identificateur unique de chaque enregistrement, deux enregistrements ne peuvent jamais avoir la même valeur pour cet attribut.
3. **record_timestamp** : Représente la date et l'heure exacte dans laquelle les enregistrements sont insérés avec le fuseau horaire international.
4. **fields.taux_dinsertion** : Représente le taux d'insertion relative à chaque enregistrement.
5. **fields.disciplines** : Représente le domaine pour lequel les données sont collectées cet attribut comporte des valeurs précises concernant le domaine (exemple Droit, économie et gestion>Droit>Sciences juridiques).
6. **fields.code_de_la_discipline** : Cet attribut comporte un identificateur unique relatif à une discipline, plus précisément il définit un identificateur pour un domaine (exemple « disc02 » est l'identificateur unique du discipline Droit, économie et gestion>Droit>Sciences juridiques).
7. **fields.diplomes_boursiers** : Définit le nombre des boursiers qui ont un diplôme dans ce domaine.
8. **fields.salaire_net_des_emplois_a_temps_plein_1er_quartile** : Représente un salaire moyen attribué dans la 1ere quartile de l'année par filière.
9. **fields.emplois_a_temps_plein** : le nombre d'enregistrement qui ont reçu un emploi à temps plein par filière.
10. **fields.salaire_net_median_des_emplois_a_temps_plein** : Le salaire net moyen des enregistrements qui ont reçus un emploi à plein temps.
11. **fields.discipline** : Représente le domaine général de l'enregistrement.
12. **fields.code_du_domaine** : Représente l'identificateur unique du domaine.
13. **fields.salaire_net_mensuel_national_1er_quartile** : Représente le salaire net par mois du premier quartile de l'année par filière.
14. **fields.nombre_de_reponses** : Représente le nombre des réponses reçus par filière.
15. **fields.emplois_exterieurs_a_la_region_de_luniversite** : Représente le nombre de recrutements à l'extérieur de la région universitaire.
16. **fields.salaire_brut_annuel_estime** : Représente le salaire annuel estimé par filière.
17. **fields.cle_disc** : Représente la clé unique de la discipline par filière.
18. **fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile** : Représente le salaire net des emplois à temps plein dans le troisième quartile de l'année.

19. **fields.emplois_cadre_ou_professions_intermediaires** : Représente le nombre des emplois dans le cadre intermédiaire.
20. **fields.genre** : Représente le genre de l'enregistrement (exemple « femme et homme »).
21. **fields.annee** : Représente l'année pour lequel l'enregistrement est étudié.
22. **fields.taux_de_chomage_national** : Représente la valeur moyenne de chômage national par filière.
23. **fields.salaire_net_mensuel_national_3eme_quartile** : Représente le salaire net national par mois dans le troisième quartile de l'année.
24. **fields.femmes** : Représente le nombre des femmes dans cette filière.
25. **fields.code_du_secteur_disciplinaire** : Représente un code unique pour un secteur disciplinaire.
26. **fields.domaine** : Représente le nom du domaine de l'enregistrement.
27. **fields.salaire_net_mensuel_median_national** : Représente le salaire net mensuel moyen national.
28. **fields.emplois_stables** : Représente le nombre d'enregistrements qui ont des postes d'emploi stables par filière.
29. **fields.code_du_secteur_disciplinaire_sise** : Représente le code unique pour un secteur disciplinaire.
30. **fields.emplois_cadre** : Représente le nombre des associations correctes du domaine à la poste occupé après recrutement.
31. **fields.situation** : Représente le période passé avant recrutement, ce période est indiqué en mois.
32. **fields.secteur_disciplinaire** : Représente le secteur disciplinaire de l'emploi.

5. Valeurs possibles pour chaque attribut :

Pour extraire les valeurs possibles qu'un attribut peut prendre en se base sur la fonction unique dans R, pour extraire les valeurs possibles pour l'attribut « fields.annee » par exemple on fait :

- `unique (Base$fields.annee)`

Remarque : « Base » est le nom de notre base des données dans R sous format Data Frame. On fait le même traitement pour tous les attributs :

```

[85] "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011"
[97] "2012" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2011" "2011" "2012" "2012" "2012"
[109] "2012" "2012" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[121] "2012" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011"
[133] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011" "2011" "2012" "2012" "2012"
[145] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[157] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011" "2012" "2012" "2012" "2012"
[169] "2012" "2012" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[181] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[193] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[205] "2012" "2012" "2011" "2011" "2012" "2012" "2012" "2012" "2011" "2011" "2012" "2012" "2012"
[217] "2012" "2011" "2012" "2012" "2012" "2012" "2012" "2011" "2011" "2011" "2012" "2012" "2012"
[229] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011" "2011" "2011" "2012" "2012" "2012"
[241] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011"
[253] "2011" "2012" "2012" "2012" "2012" "2012" "2011" "2011" "2011" "2012" "2012" "2012" "2012"
[265] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[277] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011"
[289] "2012" "2012" "2012" "2011" "2011" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[301] "2012" "2012" "2012" "2011" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[313] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[325] "2011" "2011" "2011" "2011" "2011" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[337] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011"
[349] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[361] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2011" "2012" "2011"
[373] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
[385] "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012" "2012"
> unique(Base$fields.annee)
[1] "2012" "2011"
> |

```

Figure 6 : Valeurs possibles pour l'attribut fields.annee

Valeurs Possibles pour :

1. **Datasetid** : "fr-esr-insertion_professionnelle-master_donnees_nationales".
2. **Recordid** : Attribut unique il peut prendre 389 valeurs (exemple "85d6935928c83e0d91bce5da07382ba5fbc2cd47").
3. **record_timestamp** : "2015-12-15T15:09:10-08:00".
4. **fields.taux_dinsertion** : "86" "89" "93" "85" "87" "77" "76" "84" "80" ...
5. **fields.disciplines** : "Droit, économie et gestion>Droit>Sciences juridiques" "Droit, économie et gestion>Économie>Sciences économiques" "Droit, économie et gestion>Gestion>Sciences de gestion"...
6. **fields.code_de_la_discipline** : "disc02" "disc03" "disc04" "disc06" "disc08" "disc11" ...
7. **fields.diplomes_boursiers** : "27" "32" "22" "36" "33" "NA" "38"...
8. **fields.salaire_net_des_emplois_a_temps_plein_1er_quartile** : "1450" "1600" "1700" "1290" "1310" "1300" "1260" "1500" "1400" "1330" "1350" "1430" "1250" "ns" "1200" "1470" "1370" "1440" "1480" "1720" "1520" "nd" "1550"....
9. **fields.emplois_a_temps_plein** : "94" "97" "98" "81" "90" "79" "93" "87" "57" "86" "74" "73" "77" "76" "83" "ns" "95" "96" "91" "89" "85" "88" ...
10. **fields.salaire_net_median_des_emplois_a_temps_plein** : "1720" "1990" "2020" "1700" "1730" "1600" "1400" "1800" "1630" "1500" "1650" "1780" "ns" "1680" ...
11. **fields.discipline** : "Droit" "Économie" "Gestion" "Lettres, langues, arts" "Histoire-géographie" "Autres sciences humaines et sociales" ...
12. **fields.code_du_domaine** : "DEG" "LLA" "SHS" "STS" "MEEF" "ALL".
13. **fields.salaire_net_mensuel_national_1er_quartile** : "1460" "NA".

14. **fields.nombre_de_reponses** : "2387" "1364" "6331" "136" "89" "538" "562" NA
"216" "575" "525"...
15. **fields.emplois_exterieurs_a_la_region_de_luniversite** : "42" "44" "43" "29" "37" "47" "36"
"38" "50" "35" "41" "58" "60" "ns" "46" "40" "45" "59"...
16. **fields.salaire_brut_annuel_estime** : "26900" "31000" "31500" "26500" "25000" "21800"
"28100" "25400" "23400" "25700" "27800" "ns" "26200"...
17. **fields.cle_disc** : "disc02_01_18" "disc03_30" "disc04_30" "disc06_12_30" "disc06_08_30"
"disc08_02_18"...
18. **fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile** : "2060" "2400" "2410"
"1900" "1910" "1750" "2080" "1990" "1800" "1840" "2000" "2160" "ns" "2010" "2200"...
19. **fields.emplois_cadre_ou_professions_intermediaires** : "78" "82" "83" "72" "74" "76"
"62" "90" "66" "67" "77" "79" "69" "73" "ns" "60"...
20. **fields.genre** : "femmes et hommes" "hommes" " " "femmes".
21. **fields.annee** : "2012" "2011".
22. **fields.taux_de_chomage_national** : "10.1" "9.7".
23. **fields.salaire_net_mensuel_national_3eme_quartile** : "2230" "NA".
24. **fields.femmes** : "72" "54" "56" "82" "88" "68" NA "87" "64"...
25. **fields.code_du_secteur_disciplinaire** : "disc02_01" "disc03" "disc04" "disc06_12"
"disc06_08" "disc08_02" "disc08_01" "disc11_01" "disc09"...
26. **fields.domaine** : "Droit, économie et gestion" "Lettres, langues, arts" "Sciences humaines et
sociales"
27. **fields.salaire_net_mensuel_median_national** : "1830" "1800".
28. **fields.emplois_stables** : "59" "77" "83" "56" "69" "39" "40" "48" "49" "50" "44" "46" "71"...
29. **fields.code_du_secteur_disciplinaire_sise** : "SD36" "SD38" "SD39" "SD64" "SD20" "SD28"
"SD27" "SD29" "SD32" "SD19"...
30. **fields.emplois_cadre** : "49" "55" "57" "61" "51" "36" "46" "35" "41" "NA" "48" "38" "58"
"59"...
31. **fields.situation** : "18 mois après le diplôme" "30 mois après le diplôme".
32. **fields.secteur_disciplinaire** : "Sciences juridiques" "Sciences économiques" "Sciences de
gestion" "Pluridisciplinaire lettres, sciences du langage, arts" "Littérature générale et
comparée"...

Etape 3 : Préparation de la base de données :

Cette étape consiste à bien traiter les données manquantes et les bruits dans les enregistrements, cette opération est très utile pour obtenir une base de données plus pertinente pour l'analyse.

Pour fixer les problèmes de bruitage il faut vérifier s'il y a des données manquantes pour chaque attribut déjà listées dans l'étape précédente et traiter ce problème attributs par attribut. Il faut vérifier y a-t-il des données bruitées et comment les corriger. On doit aussi décider s'il faut faire des transformations sur certains attributs.

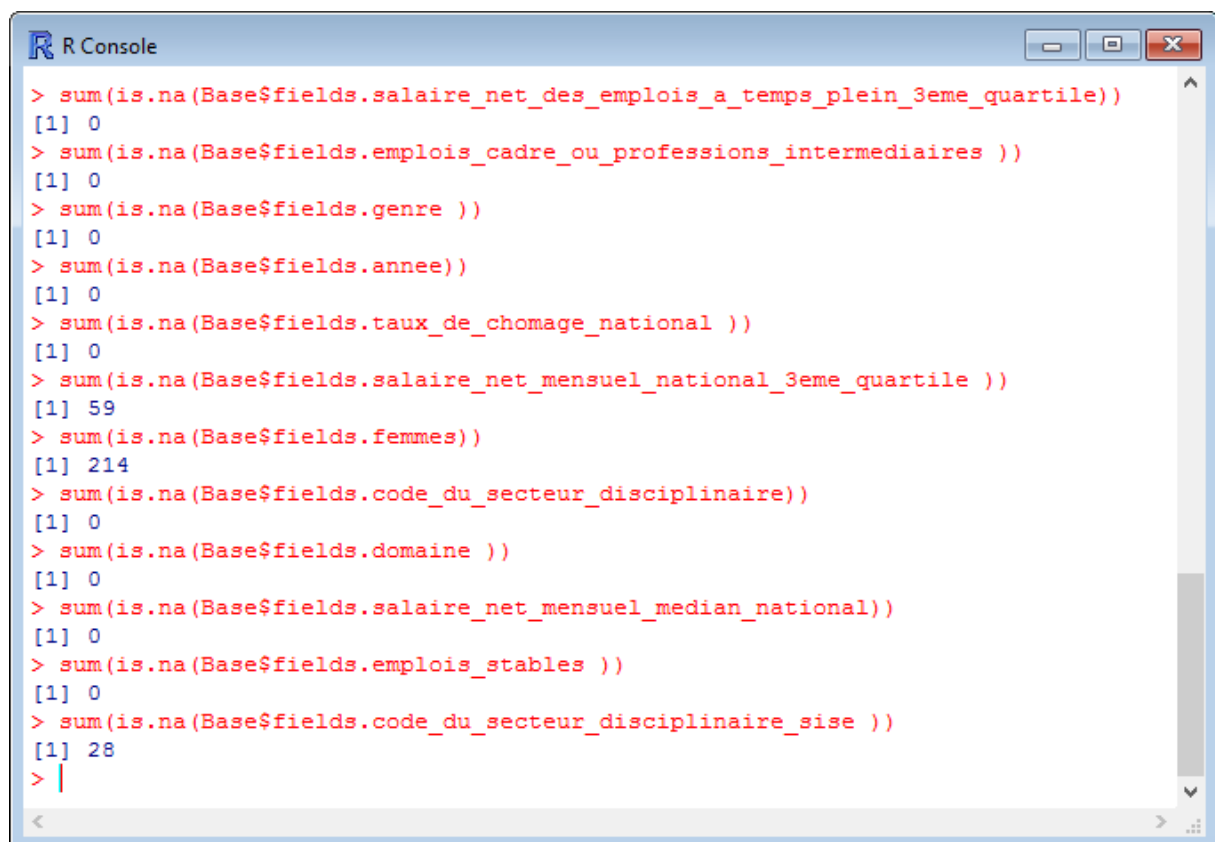
On doit procéder comme suit :

1. Vérifier s'il y a des données manquantes.
2. Comment corriger le problème ?
3. Justification de l'utilisation des transformations.

Vérification des données manquantes :

Pour vérifier s'il y a des données manquantes il faut vérifier le nombre des valeurs nuls (valeur = "" ou valeur = "NA" : "Not Available") pour tous les attributs pour cela on tape par exemple : **sum (is.na (Base\$fields.diplomes_))**

Cette ligne de commande donne le nombre des valeurs vides pour l'attribut : fields.diplomes_.



```
> sum(is.na(Base$fields.salaire_net_des_emplois_a_temps_plein_3eme_quartile))
[1] 0
> sum(is.na(Base$fields.emplois_cadre_ou_professions_intermediaires ))
[1] 0
> sum(is.na(Base$fields.genre ))
[1] 0
> sum(is.na(Base$fields.annee))
[1] 0
> sum(is.na(Base$fields.taux_de_chomage_national ))
[1] 0
> sum(is.na(Base$fields.salaire_net_mensuel_national_3eme_quartile ))
[1] 59
> sum(is.na(Base$fields.femmes))
[1] 214
> sum(is.na(Base$fields.code_du_secteur_disciplinaire))
[1] 0
> sum(is.na(Base$fields.domaine ))
[1] 0
> sum(is.na(Base$fields.salaire_net_mensuel_median_national))
[1] 0
> sum(is.na(Base$fields.emplois_stables ))
[1] 0
> sum(is.na(Base$fields.code_du_secteur_disciplinaire_sise ))
[1] 28
> |
```

Figure 7 : Extraction de nombre des valeurs nul des attributs

On fait le parcourt de toute la base attribut par attribut on obtient :

- **fields.diplomes_boursiers** : 214 Valeurs (manque majeur des données 55%).
- **fields.salaire_net_mensuel_national_1er_quartile** : 59 Valeurs (15.1%).
- **fields.nombre_de_reponses** : 214 Valeurs (manque majeur des données 55%).
- **fields.salaire_net_mensuel_national_3eme_quartile** : 59 Valeurs (15.1%).
- **fields.femmes** : 214 Valeurs (manque majeur des données 55%).
- **fields.code_du_secteur_disciplinaire_sise** : 28 Valeurs (champ non pertinent 7.19%).
- **fields.emplois_cadre** : 107 Valeurs (27.5%).

Initialement la base de données est composées par des chaines de caractères donc on ne peut pas faire la détection des valeurs bruitées, aussi les données sont gravement manquantes cela peut influencer les résultats de l'analyse ultérieurement.

Propositions :

On doit changer tous les chaines de caractères qui portent des valeurs numériques càd on doit remplacer la valeur "42" de l'attribut **fields.emplois_exterieurs_a_la_region_de_luniversite** en une valeur numérique **42** en conservent les valeurs manquantes initialement.

La méthode `strtoi(x)` en R permet de transformer une chaine de caractère en une valeur numérique et de préserver les valeurs nulles.

`strtoi (Base$fields.diplomes_boursiers)` est un exemple de transformation des valeurs de l'attribut **fields.diplomes_boursiers** en des valeurs numériques.

La figure 8 ci-dessous représente la liste des valeurs de l'attribut **fields.diplomes_boursiers** avant transformation :

```

R Console
> Base$fields.diplomes_boursiers
[1] "27" "32" "22" "22" "36" "33" "36" NA NA NA NA NA NA "38" "29"
[16] "33" "29" NA NA NA NA NA NA "29" "27" "40" "37" "29" NA NA
[31] NA "29" "29" "18" "ns" "29" "29" NA NA NA NA "29" NA NA NA
[46] "29" NA NA NA NA NA NA "28" "35" "29" NA NA NA NA NA NA
[61] NA "35" "35" "29" "29" NA NA NA NA NA "32" "27" "29" "29"
[76] "29" "29" NA NA NA NA "26" "23" "29" "29" NA NA NA NA NA
[91] NA NA "29" "29" "ns" "29" "34" "29" "34" "34" "46" "41" "36" "29" "29"
[106] NA NA NA "28" "18" "29" "29" NA NA NA NA NA NA "40" "40"
[121] "ns" "29" NA NA NA NA NA NA "31" "18" "18" "29" NA NA NA
[136] NA NA NA "ns" "29" "29" NA NA NA NA NA NA NA NA NA
[151] NA NA NA NA NA NA NA NA "26" "28" "23" "17" "38" "29" NA
[166] NA "25" "25" "24" "32" "29" NA NA NA "22" NA NA NA NA NA
[181] NA NA NA NA "36" "36" "32" "28" "ns" NA NA NA NA NA "41"
[196] NA NA NA NA NA NA "26" "ns" "30" "30" "29" "29" NA NA
[211] NA NA "29" "38" NA NA NA "29" NA NA NA "ns" "29" "29" "29"
[226] "29" NA NA NA NA NA "27" "31" "32" "29" "29" "29" NA NA NA
[241] "33" "36" "35" "32" NA NA NA NA NA "24" "29" "29" NA "29"
[256] "30" "30" "29" "29" NA NA NA NA "37" "22" "31" "31" "32" "ns"
[271] NA NA NA NA "32" "27" "35" "29" NA NA NA NA NA NA "29"
[286] "34" "17" "29" NA NA NA "29" "29" "29" NA NA NA NA NA NA
[301] NA "38" "31" "29" "29" NA NA NA NA NA "23" "23" "23" NA
[316] NA NA "26" "26" "27" "ns" "33" "34" NA "29" "29" "29" "29" NA
[331] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[346] "26" NA "29" NA "31" "35" "30" "32" NA NA NA NA NA NA "23"

```

Figure 8 : fields.diplomes_boursiers avant transformation

La figure 9 ci-dessous représente le résultat qu'on obtient après transformation avec **strtoi** (**Base\$fields.diplomes_boursiers**) :

```

R Console
[271] NA NA NA NA "32" "27" "35" "29" NA NA NA NA NA NA "29"
[286] "34" "17" "29" NA NA NA "29" "29" "29" NA NA NA NA NA NA
[301] NA "38" "31" "29" "29" NA NA NA NA NA NA "23" "23" "23" NA
[316] NA NA "26" "26" "27" "ns" "33" "34" NA "29" "29" "29" "29" NA
[331] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[346] "26" NA "29" NA "31" "35" "30" "32" NA NA NA NA NA NA "23"
[361] "29" "30" "ns" NA NA NA NA NA "29" NA "29" NA NA NA
[376] NA NA NA NA "34" "40" "26" "32" "46" "35" "35" "26" NA NA
> strtoi(Base$fields.diplomes_boursiers)
[1] 27 32 22 22 36 33 36 NA NA NA NA NA NA 38 29 33 29 NA NA NA NA NA NA 29 27
[26] 40 37 29 NA NA NA 29 29 18 NA 29 29 NA NA NA NA 29 NA NA NA 29 NA NA NA NA
[51] NA 28 35 29 NA NA NA NA NA NA NA 35 35 29 29 NA NA NA NA NA NA 32 27 29 29
[76] 29 29 NA NA NA NA NA 26 23 29 29 NA NA NA NA NA NA NA 29 29 NA 29 34 29 34 34
[101] 46 41 36 29 29 NA NA NA 28 18 29 29 NA NA NA NA NA NA 40 40 NA 29 NA NA NA
[126] NA NA NA 31 18 18 29 NA NA NA NA NA NA NA 29 29 NA NA NA NA NA NA NA NA
[151] NA NA NA NA NA NA NA NA NA 26 28 23 17 38 29 NA NA 25 25 24 32 29 NA NA NA 22
[176] NA NA NA NA NA NA NA NA NA 36 36 32 28 NA NA NA NA NA NA 41 NA NA NA NA NA
[201] NA NA 26 NA 30 30 29 NA NA NA NA NA 29 38 NA NA NA 29 NA NA NA 29 29 29
[226] 29 NA NA NA NA NA 27 31 32 29 29 29 NA NA NA 33 36 35 32 NA NA NA NA NA NA
[251] 24 29 29 NA 29 30 30 29 29 29 NA NA NA NA 37 22 31 31 32 NA NA NA NA NA 32
[276] 27 35 29 NA NA NA NA NA NA 29 34 17 29 NA NA NA 29 29 29 NA NA NA NA NA NA
[301] NA 38 31 29 29 NA NA NA NA NA NA 23 23 23 NA NA NA 26 26 27 NA 33 34 NA 29
[326] 29 29 29 29 NA NA NA NA NA NA NA NA NA NA NA NA NA NA 26 NA 29 NA 31
[351] 35 30 32 NA NA NA NA NA NA 23 29 30 NA NA NA NA NA NA 29 NA 29 NA NA NA
[376] NA NA NA NA 34 40 26 32 46 35 35 26 NA NA

```

Figure 9 : Résultat de la transformation de fields.diplomes_boursiers

Après cette transformation on peut bien détecter les valeurs manquantes car ils ne sont pas des valeurs numériques.

Pour le cas des valeurs manquantes ou bruitées il faut appliquer des techniques pour bien remplacer ces valeurs, en cours on a étudiés divers techniques tels que :

- Les ignorées.
- Remplir les valeurs manquantes manuellement.
- Remplir les valeurs manquantes automatiquement par :
 - Constante globale : « unknown » par exemple, créer une nouvelle classe...
 - Valeur moyenne.
 - Valeur moyenne des attributs de la même classe.
 - La valeur la plus probable.

La solution la plus adaptable est de traiter chaque attribut à part car on doit faire attention aux infections qui peuvent se produire lors de calcul de dissimilarité...

En général on va recours à la technique de remplissage manuelle des valeurs manquantes car la base ne contient pas un nombre très élevé des enregistrements de plus les attributs qui possède des valeurs manquantes ne sont pas nombreux.

Le remplissage des valeurs manquantes pour les attributs pertinentes sera par le remplissage avec des constantes globales dans notre exemple on va les remplacer par -1 puisque après une examination détaillé de la base et des valeurs des attributs on remarque que le remplissage des valeurs manquantes avec des autres valeurs peut entrainer des sérieuses problèmes ultérieurement.

Le remplissage des valeurs manquantes avec une constante globale peut nous faciliter le calcul.

Après le remplissage de la base il nous reste que de traiter les valeurs bruitées.

Détection de bruitage « Outliers » :

Pour la détection des « Outliers » on fait le résumé de toutes les variables de la base « summary (Base) » pour explorer visuellement si on a des valeurs mal placées (Peut être une source des « Outliers »). Après la visualisation du résumé des variables on remarque que certains attributs peuvent présenter des « Outliers ».

On choisit 3 attributs parmi eux pour la démonstration :

- **fields.salaire_net_median_des_emplois_a_temps_plein.**
- **fields.nombre_de_reponses.**
- **fields.salaire_net_mensuel_national_1er_quartile.**

La figure 10 ci-dessous représente le résumé de ces 3 attributs :

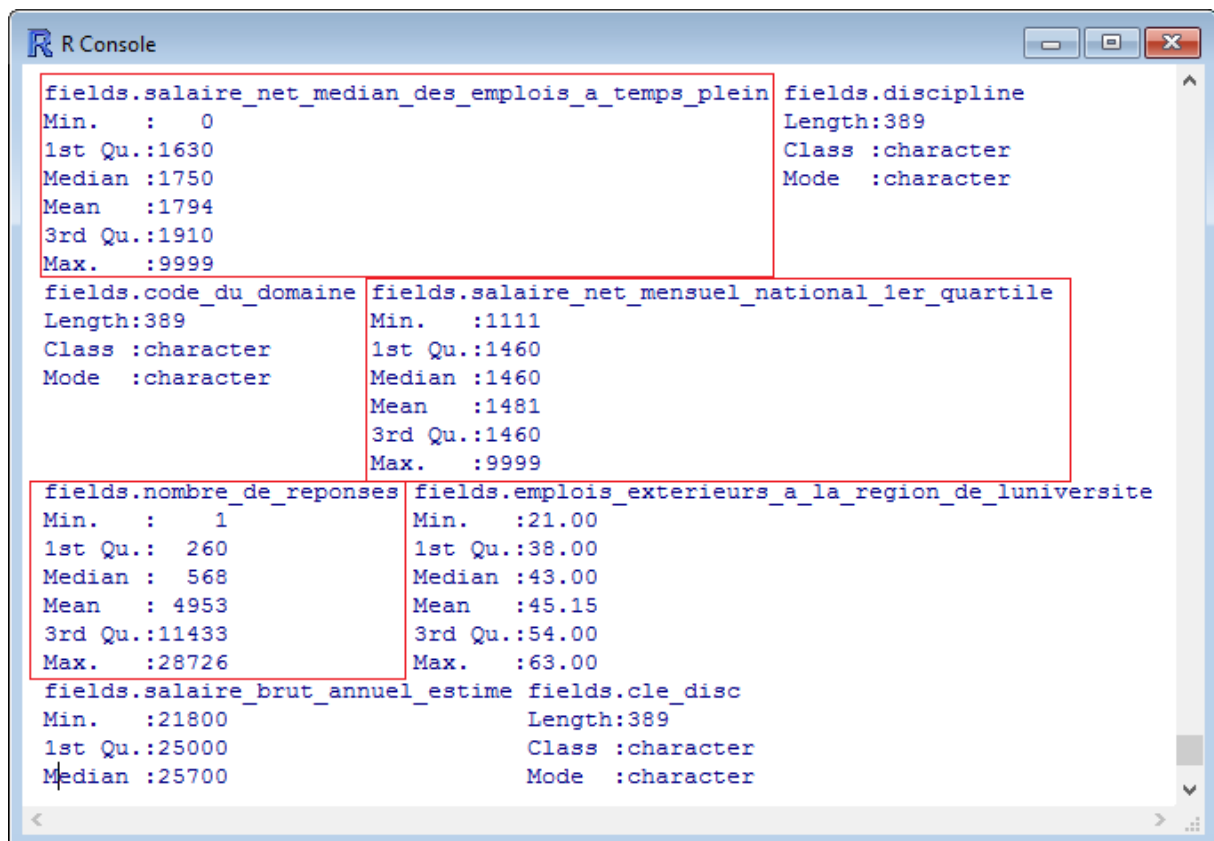


Figure 10 : Détection des "Outliers"

On remarque par exemple pour l'attribut **fields.salaire_net_mensuel_national_1er_quartile** que la valeur minimale que cet attribut peut prendre est 1111 et une valeur moyenne de 1460 et pour le 3^{ème} quartile on trouve encore une valeur de 1460 mais pour la valeur maximale on trouve 9999 ce qui signifie qu'il existe des bruitages « Outliers ».

Même chose pour les deux autres attributs, la visualisation du résumé de la base peut nous donner une idée sur les données bruitées mais pour les étudier il faut recourir à l'Exploration monodimensionnelle et bi-dimensionnelle. Dans les étapes suivantes on va utiliser les histogrammes, les « Boxplots » et les « Quatile Plot » pour détecter les données bruitées graphiquement pour les corriger ultérieurement à l'aide des techniques de traitement des données bruitées étudiées en cours.

Comment Corriger les données bruitées ?

Il existe divers techniques de correction des données bruitées tels que :

- Binning
- Regression
- Clustering
- Combined computer and human inspections...

On finit l'étape de préparation de la base de données par la réponse à la question : **est-ce que ce traitement des « Outliers » est important ?**

Réponse : Absolument oui, le traitement des données bruitées nous garantit une idée globale sur le contenu de la base, de plus cette étude nous permet de bien analyser les données pour faciliter l'extraction des connaissances.

Etape 4 : Exploration mono-dimensionnelle

Cette étape consiste à visualiser les attributs à l'aide des histogrammes, des « Boxplots » et les « Quantile Plot ».

Dans cette étape et à cause de nombre élevé des attributs on va faire le traitement de 3 attributs uniquement pour l'exploration mono-dimensionnelle.

La suite de cette étape consiste à étudier les attributs :

- **fields.salaire_net_median_des_emplois_a_temps_plein.**
- **fields.nombre_de_reponses.**
- **fields.salaire_net_mensuel_national_1er_quartile.**

1. Histogrammes :

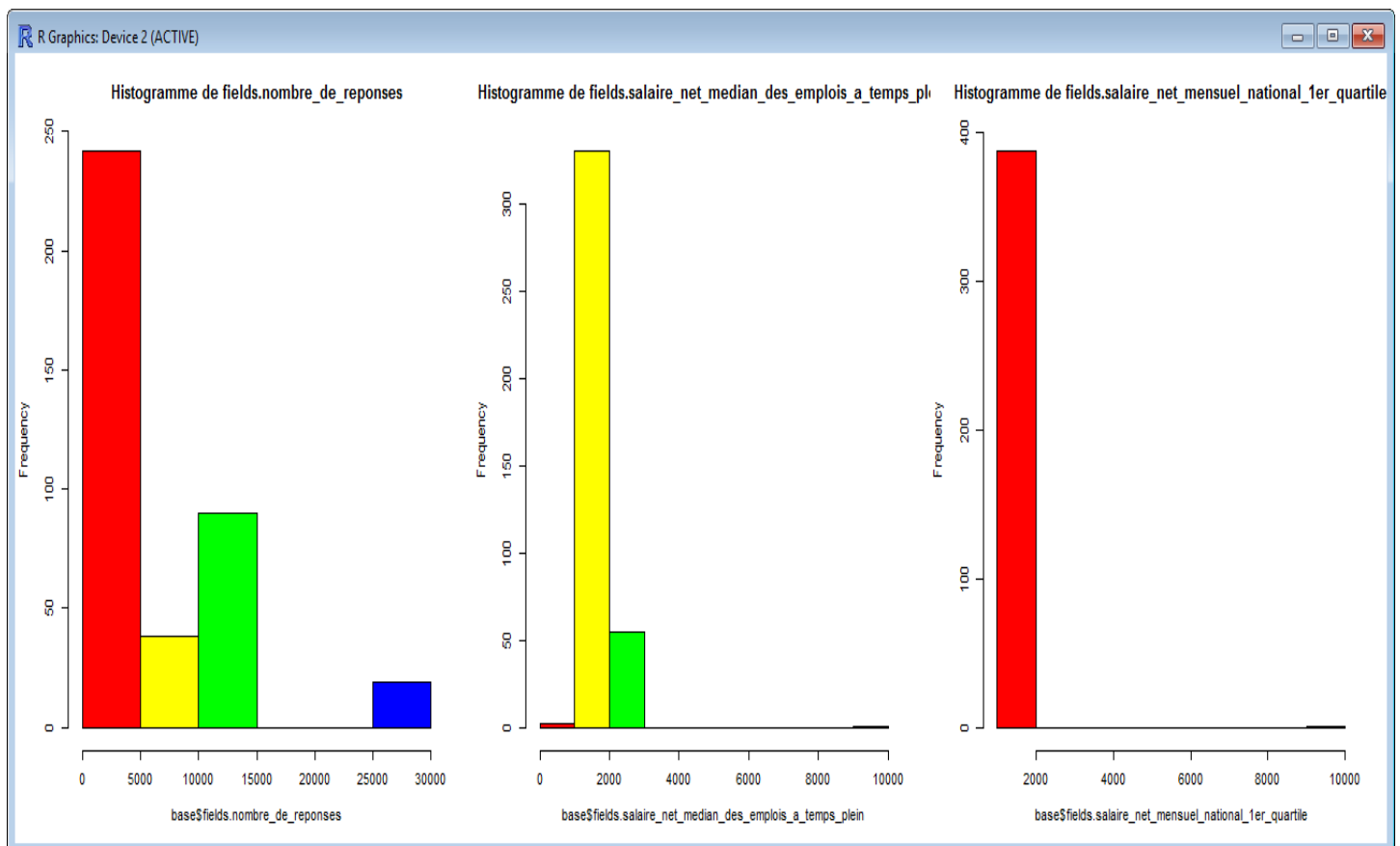


Figure 11 : Histogrammes des trois attributs

Une première impression des histogrammes des trois attributs qu'il existe des données bruitées. On remarque que l'histogramme de l'attribut **fields.salaire_net_median_des_emplois_a_temps_plein** montre clairement des données bruitées dans l'intervalle inférieure à 1000 et dans l'intervalle entre 900 et 1000, ces valeurs sont de faible fréquence ce qui montre qu'ils sont des « Outliers ».

Aussi pour l'attribut **fields.salaire_net_mensuel_national_1er_quartile** les valeurs de hautes fréquences sont dans l'intervalle entre 0 et 200 mais on trouve des valeurs de faible fréquence dans l'intervalle 9000 et 10000, ces derniers semblent d'être des « Outliers ».

2. Les « Boxplots » :

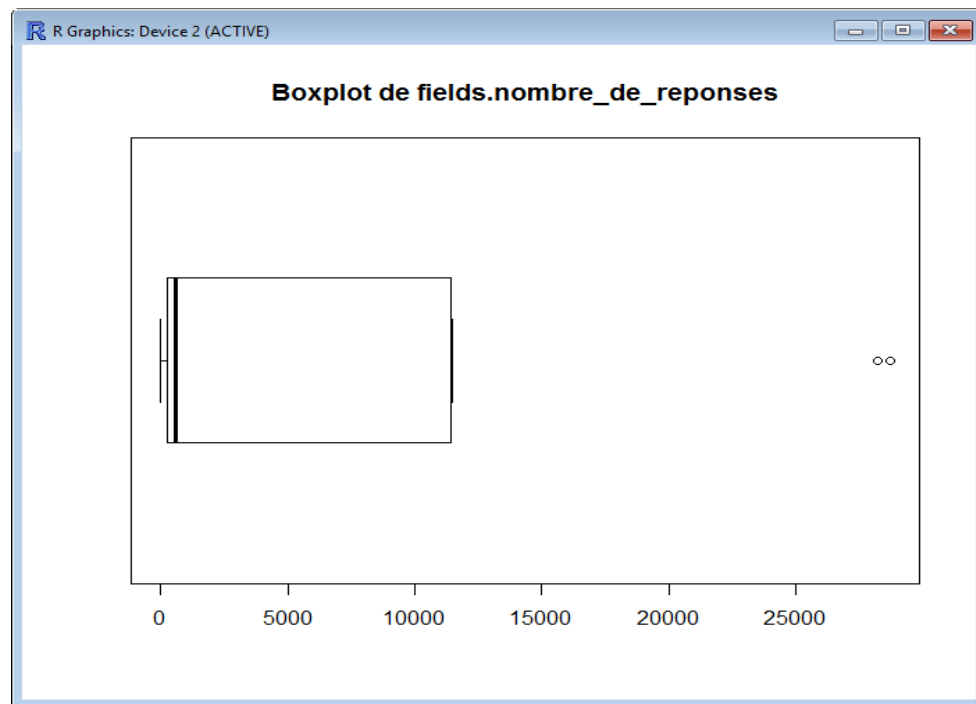


Figure 12 : "Boxplot" de l'attribut *fields.nombre_de_reponses*

La Figure 12 ci-dessus représente les « Outliers » présents dans les valeurs de l'attribut **fields.nombre_de_reponses**, on constate que les valeurs des données bruitées sont supérieures à 25000, Les valeurs sont généralement réparties entre 0 et 13000. On remarque que l'histogramme donne toujours des meilleurs résultats que les « Boxplots ».

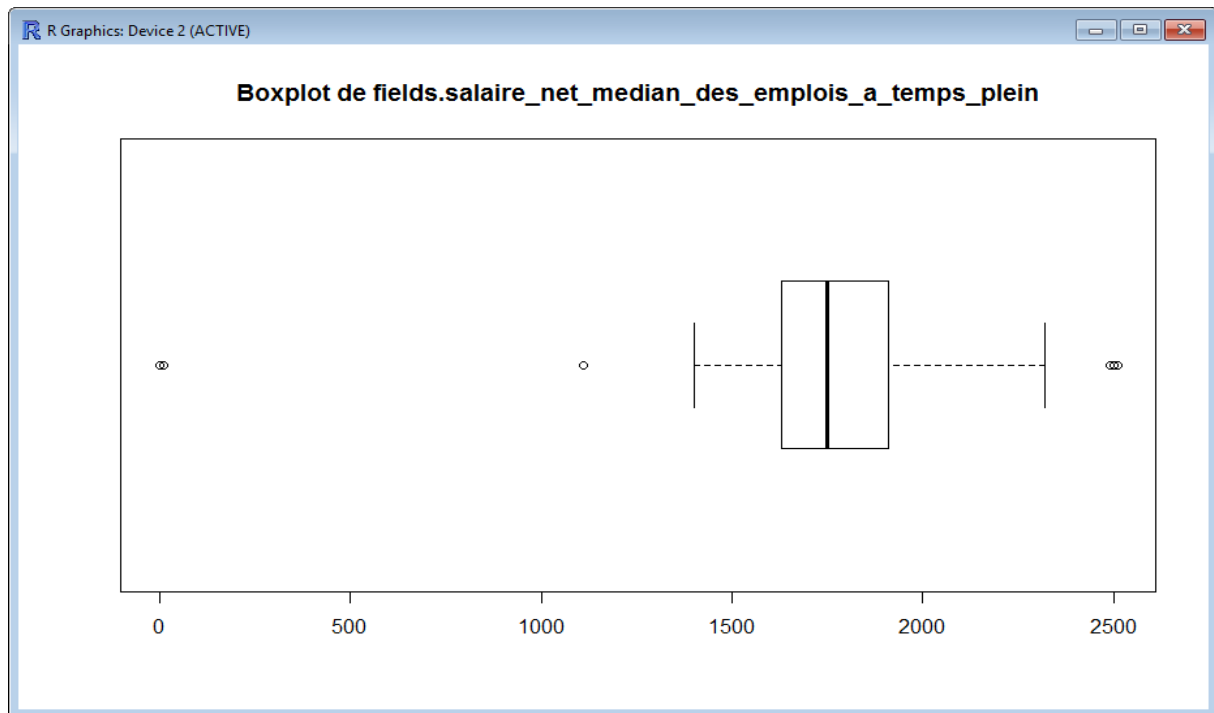


Figure 13 : "Boxplot" de l'attribut `fields.salaire_net_median_des_emplois_a_temps_plein`

La Figure 13 ci-dessus représente les « Outliers » présents dans les valeurs de l'attribut `fields.salaire_net_median_des_emplois_a_temps_plein`, on constate que les valeurs des données bruitées sont réparties en divers intervalles (valeurs proches à 0, entre 1000 et 1200 et entre 2400 et 2500), Les valeurs sont généralement réparties entre 1500 et 2000.

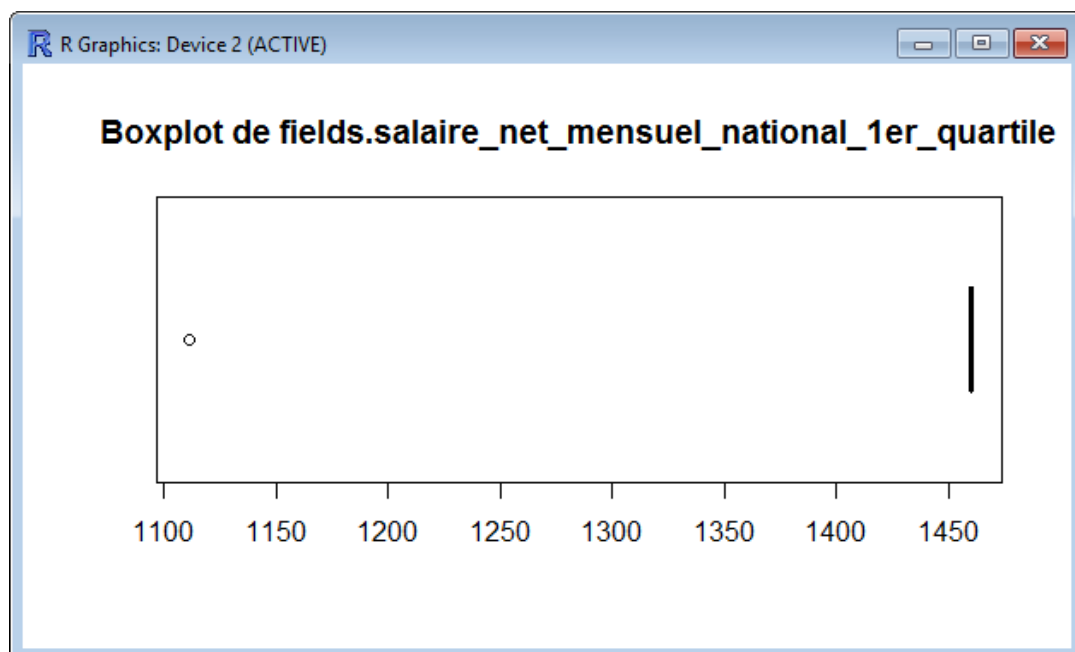


Figure 14 : "Boxplot" de l'attribut `fields.salaire_net_mensuel_national_1er_quartile`

La Figure 14 ci-dessus représente les « Outliers » présents dans les valeurs de l'attribut **fields.salaire_net_mensuel_national_1er_quartile**, on constate que les valeurs des données bruitées sont entre 1100 et 1150, Les valeurs sont normalement supérieures à 1450.

Comme la montre les figures 12,13 et 14 les « Boxplots » des trois attributs montrent aussi les « Outliers ».

3. Les « Barplots » :

Les « Barplots » permet la représentation d'un attribut qualitatif, pour cela on doit tout d'abord construire le table de contingence à l'aide de fonction `table ()`.

La figure ci-dessous représente les « Barplots » des 3 attributs :

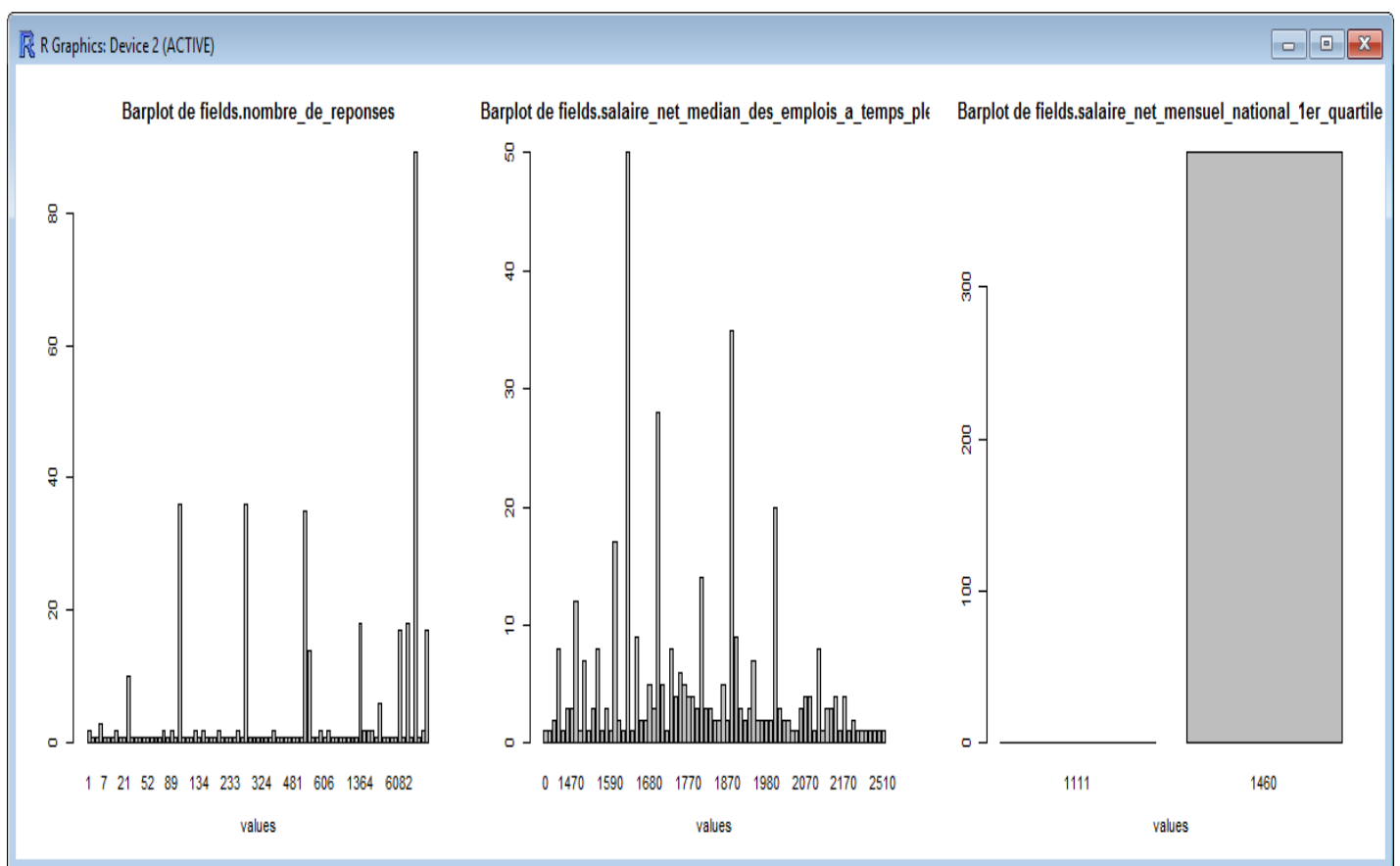


Figure 15 : "Barplots" des trois attributs

Les « Barplots » nous aide à bien voir la distribution des valeurs sur un repère, cette opération est utile pour visualiser les valeurs que l'attribut peut prendre et détecter les « Outliers » visuellement.

Etape 5 : Exploration bi-dimensionnelle des données

Cette étape consiste à représenter des couples d'attribut dans une même figure, cette opération facilite la comparaison des couples d'attribut selon la représentation graphique.

Pour effectuer la comparaison des couples d'attribut on choisit les attributs :

- **fields.salaire_net_median_des_emplois_a_temps_plein.**
- **fields.nombre_de_reponses.**
- **fields.annee.**
- **fields.salaire_brut_annuel_estime.**

Pour ces attributs on doit comparer leurs valeurs graphiquement en des Histogrammes, Boxplot et Pie...

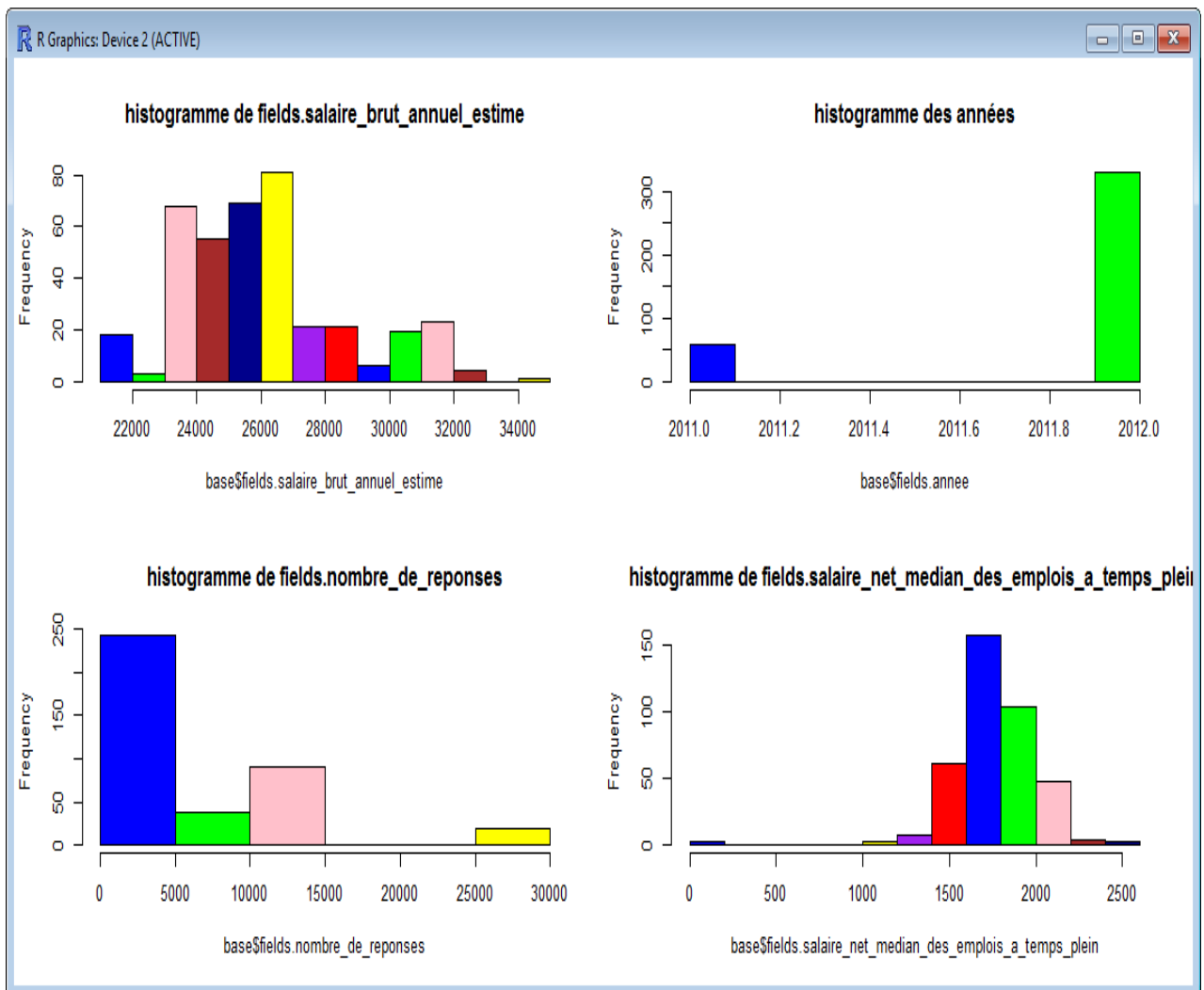


Figure 16 : Histogrammes des quatres attributs

On peut comparer les valeurs de chaque attribut selon les valeurs distribuées sur le repère, on peut aussi afficher les graphes ou encore charte graphique « Pie » pour une visualisation plus simple mais il faut transformer la classe de l'attribut en facteur « Factor », cette étape est réalisé en R à l'aide de la ligne de commande : **as.factor ()**.

Exemples de démonstration : les attributs **fields.taux_de_chomage_national** et **fields.annee** :

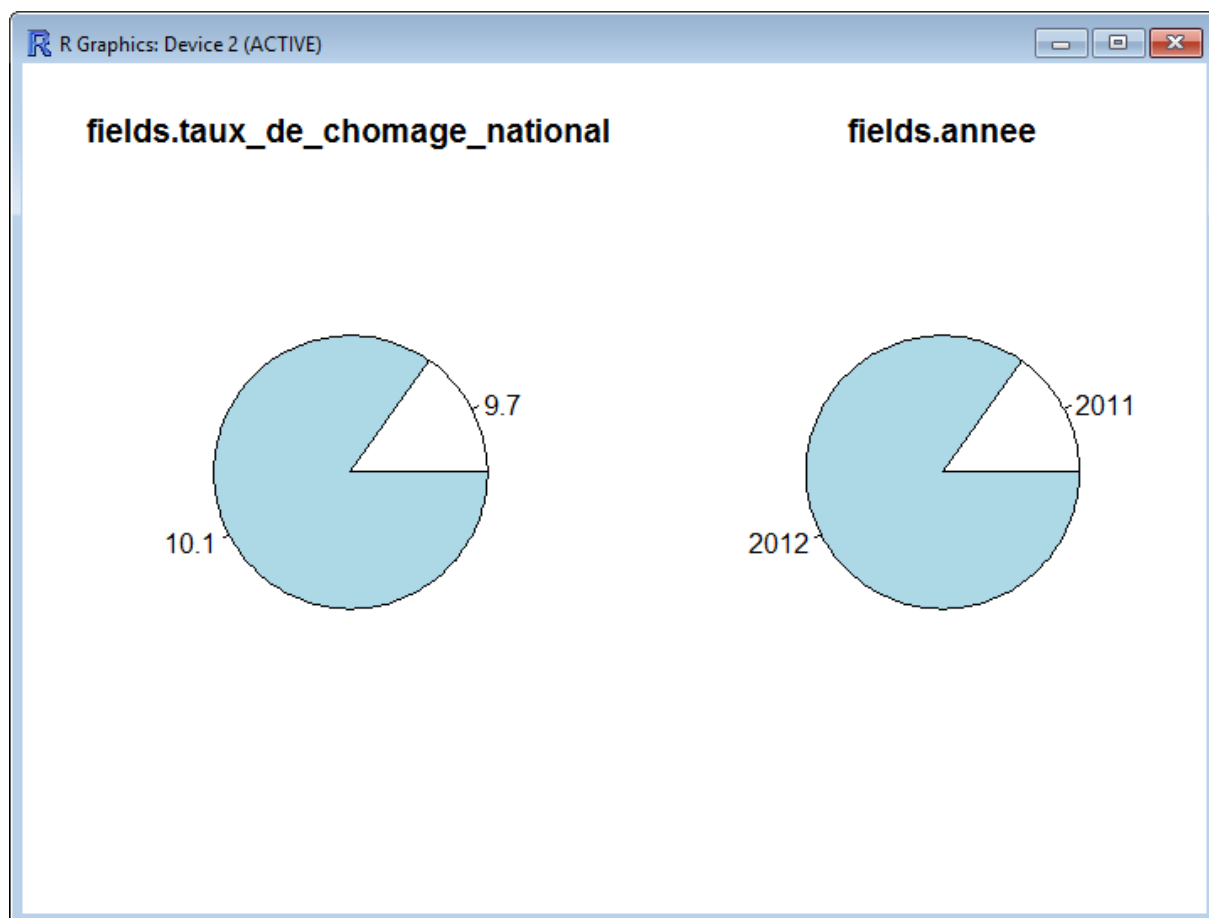


Figure 17 : Affichage des "Pie Chart" des deux attributs

L'outil R présente aussi la possibilité de l'exploration des valeurs à l'aide des nuages de points, pour réaliser cette exploration on va utiliser les attributs :

- **fields.nombre_de_reponses**
- **fields.annee**

La figure ci-dessous représente le nombre des réponses réparties selon la classe **fields.annee**, le couleur rouge représente les valeurs de **fields.nombre_de_reponses** relative à l'année 2011.

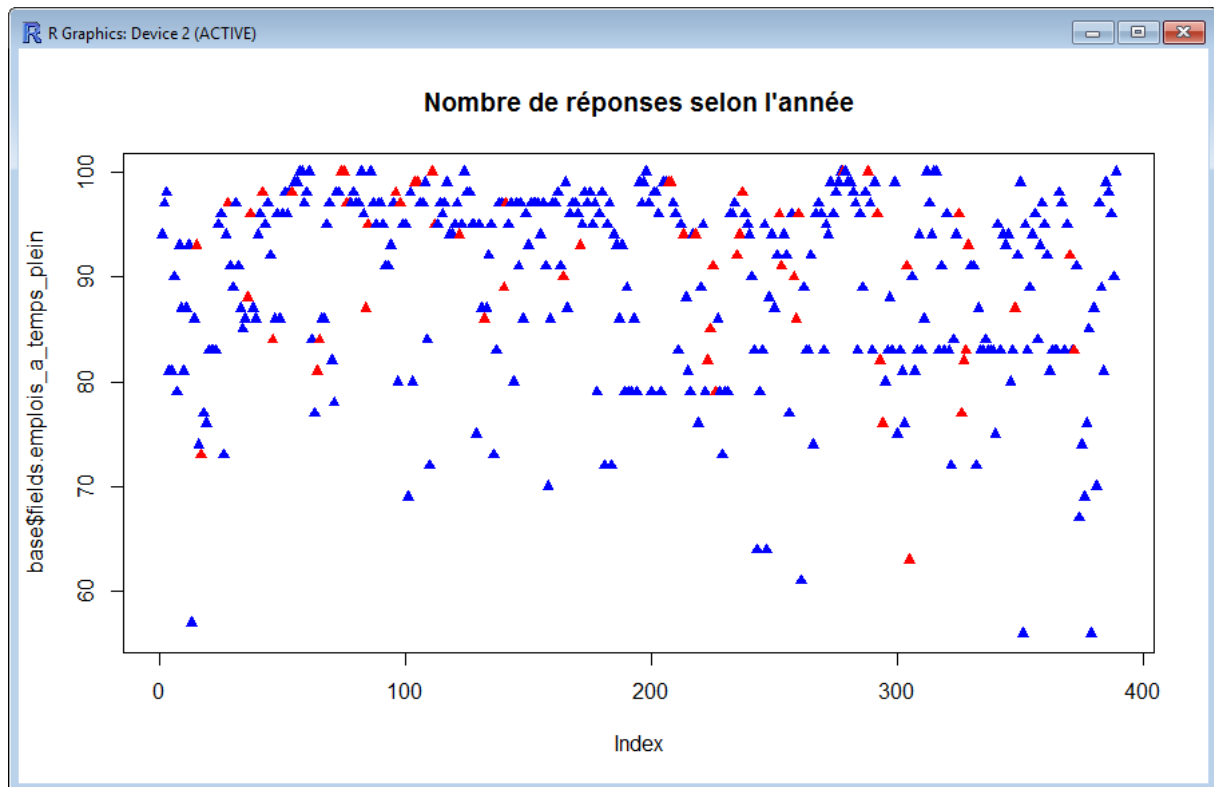


Figure 18 : Représentation de nombre de réponses selon l'année

On a fini la partie de l'importation et de l'exploration des données issue de notre base de données, les étapes suivantes consistent à manipuler le bruitage dans les données, il existe divers techniques qu'on peut les utilisées pour fixer ces problème mais le choix reste très important puisque ceci peut influencer le calcul de l'analyste ultérieurement.